ATTENTION-BASED ATROUS CONVOLUTIONAL NEURAL NETWORKS: VISUALISATION AND UNDERSTANDING PERSPECTIVES OF ACOUSTIC SCENES

Zhao Ren[1], *Qiuqiang Kong*[2], *Jing Han*[1], *Mark D. Plumbley*[2], *Björn W. Schuller*[1,3]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[3] GLAM – Group on Language, Audio & Music, Imperial College London, UK

zhao.ren@informatik.uni-augsburg.de

ABSTRACT

The goal of Acoustic Scene Classification (ASC) is to recognise the environment in which an audio waveform has been recorded. Recently, deep neural networks have been applied to ASC and have achieved state-of-the-art performance. However, few works have investigated how to visualise and understand what a neural network has learnt from acoustic scenes. Previous work applied local pooling after each convolutional layer, therefore reduced the size of the feature maps. In this paper, we suggest that local pooling is not necessary, but the size of the receptive field is important. We apply atrous Convolutional Neural Networks (CNNs) with global attention pooling as the classification model. The internal feature maps of the attention model can be visualised and explained. On the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 dataset, our proposed method achieves an accuracy of 72.7 %, significantly outperforming the CNNs without dilation at 60.4 %. Furthermore, our results demonstrate that the learnt feature maps contain rich information on acoustic scenes in the time-frequency domain.

Index Terms— deep neural networks, atrous convolutional neural networks, attention pooling, acoustic scene classification

1. INTRODUCTION

To recognise acoustic environments automatically, Acoustic Scene Classification (ASC) [1] has been a main objective of research in computer audition [2, 3]. It aims at classifying acoustic scenes through computational algorithms including signal processing and machine learning. A variety of applications could benefit from ASC, including mobile robots [4], context-aware computing [5], and wearable devices [6].

Previous methods applied Support Vector Machines (SVMs) [7] and Hidden Markov Machines (HMMs) [8] to ASC. Recently, neural network based methods including fully connected neural networks [9], Recurrent Neural Networks (RNNs) [10, 11], and Convolutional Neural Networks (CNNs) [11], have achieved the state-of-the-art performance in ASC. Neural networks are effective at extracting high-level features to classify unseen data. However, previous work for audio classification [12] did not visualise and analyse the internal layers of CNNs.

This paper aims to visualise high-level representations in CNNs. For example, while spectrogram images of audio waveforms are the input, the time-frequency units in a feature map can be localised according to their contribution. This idea of localisation is inspired from image-based object localisation [13]. Our work can help better explain which time-frequency components contribute to ASC and can be further used for sound segmentation or separation [14].

There are two difficulties in visualising high-level representations in CNNs. Firstly, the learnt representations depend on global pooling after the last convolutional layer. Global max or average pooling result in accurate classification, but tend to under- or overestimate the units in feature maps [14]. Global attention pooling has been proposed to adaptively attend to the units [15, 16]. However, in [16], the learnt low resolution representations lost the time-frequency details due to stride convolution, which is similar with local pooling layers.

In this paper, we discover that local pooling is not necessary, but the size of the receptive field is important for ASC. We propose to use atrous CNNs [17] with a large receptive field instead of local pooling to fix the size of feature maps. Then, a global attention pooling layer is applied on the feature maps to learn the time-frequency units' contributions.

2. RELATED WORK

Our proposed attention-based atrous CNNs build on previous work using attention-based CNNs [16]. In that work, we extracted attention matrices with a size of 4×20 and applied a basic analysis. However, the resulting low resolution feature maps could not describe the time-frequency properties of acoustic scenes in detail.

To fix the size of the feature maps at each convolutional layer, the simplest solution is a vanilla CNN model without local pooling layers. However, this increases both time and

space complexities and results in sub-optimisation. Encoder-decoder CNNs were proposed in [18], employing a decoder to up-sample the feature maps using transferred pool indices from the encoder. Similarly, in [19], Fully Convolutional Networks (FCNs) used deconvolutional layers to up-sample the feature maps. However, both encoder-decoder CNNs and FCNs are comprised of pooling and up-sampling layers, therefore require strongly labelled data for pixel-wise classification. Datasets in ASC can be considered weakly labelled as only one acoustic class is annotated for each audio wave. In a separate study [17], atrous CNNs were used with a dilated receptive field instead of pooling and up-sampling, obtaining state-of-art results for the task of semantic image segmentation. Motivated by this success, we herein use atrous CNNs, with a weakly labelled ASC dataset and combined with an attention model to improve the visualisation of representations.

3. METHODOLOGY

3.1. Baseline CNNs

CNNs have been successfully used for tasks of audio classification [12, 20, 21]. In our work, log mel spectrogram images [22] are extracted from audio waveforms as the input of CNNs. The baseline CNN model consists of four convolutional layers. Low-level convolutional layers are designed to extract low-level features; high-level convolutional layers are good at learning more abstract representations such as acoustic sounds patterns [23]. A local max pooling operation with a kernel size of 2×2 is applied after each convolutional layer to extract the shift-invariant features [24] (Fig. 1 (a)). Then, a global pooling layer [12] is applied to the final feature maps. Finally, a softmax non-linearity is utilised to predict the probabilities of scene classes.

3.2. Atrous CNNs

However, a local max pooling operation in baseline CNNs results in feature maps with a small size (Fig. 1 (a)). Therefore, the feature maps cannot be pixel-wisely mapped to spectrogram images. The simplest solution is to remove all local max pooling layers so that the size of the feature maps is fixed (Fig. 1 (b)). In the experiment section, we will show that the CNNs in Fig. 1 (b) underperform the baseline CNNs.

Interestingly, we discover that this underperformance is not caused by removing local max pooling layers. Instead, it arises from the reduced size of receptive field relative to the input of CNNs. The size of a receptive field is *number of frequency bins* $\times$ *number of time frames* during the convolution operation. Without local max pooling, the size of a receptive field increases linearly with the number of layers; with local max pooling, it increases exponentially with the number of layers.

We introduce atrous CNNs [25] to improve the performance without local max pooling. Atrous CNNs have been applied to high resolution image segmentation [17] and audio generation [25] by fixing the size of feature maps. Atrous CNNs use dilated convolutional kernels (Fig. 1 (c)); therefore, the size of the receptive field increases exponentially with the number of layers. The dilated convolutional kernel is a sparse kernel so that the number of parameters does not increase compared to the baseline CNNs.

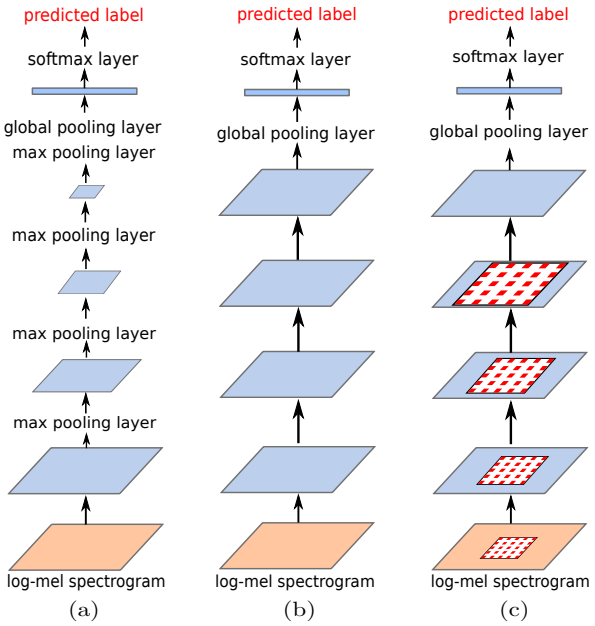

Fig. 1. Three CNNs: (a) Baseline CNNs, (b) CNNs without local max pooling layers, (c) Atrous CNNs.

3.3. Pooling Mechanism

Each feature map at the final convolutional layer has a size of $C \times F \times T$, where C, F, and T denote the number of channels, number of frequency bins, and number of time frames, respectively. Global pooling includes max [12], average [26], and attention pooling [16]. Then, a fully connected layer is applied to the output of global pooling to predict the probability of each class. Global max or average pooling has the drawback of under- or overestimating the units in feature maps. On the other hand, attention pooling can adaptively learn the contributions of the time-frequency units. Attention pooling consists of an attention and a classification branch,

$$P_{ft} = A_{ft} / \sum_{f=1}^{F} \sum_{t=1}^{T} A_{ft}, \tag{1}$$

$$Y = \sum_{f=1}^{F} \sum_{t=1}^{T} P_{ft} \cdot C_{ft}, \tag{2}$$

where A, P, and C each are the attention, probability, and classification matrices, and Y denotes the probabilities of classes.

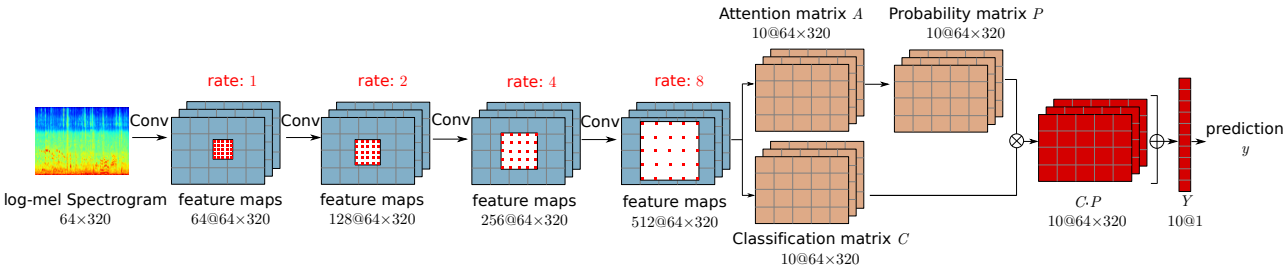

Fig. 2. The framework of our proposed attention-based atrous CNNs. The log mel spectrogram images with a size of 64×320 are fed into CNNs with four dilated convolutional layers and an global attention pooling layer. The size of the feature maps is represented as $number\,of\,channels@frequency\,bins \times time\,frames$, and the size of holes within a kernel is adapted by 'rate'.

In this paper, we additionally apply Region of Interest (ROI) pooling [27] followed by global max pooling for an experimental comparison. ROI pooling is achieved by a local max pooling operating in a 16×16 aliquoted feature map at the final convolutional layer, to bring about the same effect of the baseline CNNs using four 2×2 max pooling layers.

4. EXPERIMENTAL RESULTS

4.1. Database

Our proposed approach is evaluated on the development set of the ASC task of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 challenge [28]. The dataset contains 10 acoustic scene classes and each audio recording has a duration of 10 seconds. This ASC task consists of two subtasks defined by matching or mismatching devices.

4.2. Setup

Log mel spectrogram images with a size of 64 mel frequency bins and 320 time frames are extracted from the audio recordings with a Hamming window size of 2 048. The overlap is set to satisfy that 320 time frames are sampled in single spectrogram. We train the models for 15 000 iteration steps with a batch size of 16 to use single Graphics Processing Unit (GPU) sufficiently. The 'Adam' optimiser [29] is employed with an initial learning rate of 0.001. The learning rate is decreased by a factor of 0.9 at every 200 iteration steps to stabilise the training procedure. The set-up of the number of mel frequency bins and initial learning rate are empirical.

4.3. Results and Discussion

We apply different global poolings on the baseline CNNs, CNNs without local max pooling and atrous CNNs. Their results are shown in Table 1. To reduce the risk of overfitting caused by excessive parameters for 10-class classification, we only experiment flattening on the baseline CNNs which have feature maps with a small size of 4×20. In the baseline CNNs,

Table 1. Performance comparison of CNN topologies with flattening and five global pooling models, including max, average ('avg'), ROI, attention ('att'), and the combination of ROI and attention ('roi+att'), evaluated on two subtasks (SUBA on device *A* and SUBB on three devices *A*, *B*, and *C*) of accuracy.

Accuracy		**SUBA**	**SUBB**		
Network	Pooling	*A*	*A*	*B*	*C*
Baseline CNN	flatten	.609	.616	.494	.467
Baseline CNN	max	.686	.698	.572	.578
Baseline CNN	avg	.691	.658	.572	.578
Baseline CNN	att	.724	.726	.622	.561
CNN w/o local pool	max	.604	.619	.467	.522
CNN w/o local pool	avg	.628	.591	.544	.500
CNN w/o local pool	roi	.616	.617	.506	.439
CNN w/o local pool	att	.621	.596	.450	.433
CNN w/o local pool	roi+att	.681	.692	.561	.506
Atrous CNN	max	.688	.697	.600	.594
Atrous CNN	avg	.691	.672	.628	.600
Atrous CNN	roi	.652	.626	.483	.439
Atrous CNN	att	**.727**	**.732**	**.644**	**.622**
Atrous CNN	roi+att	.726	.722	.572	.567

the global max, average, and attention pooling outperform flattening on both subtasks. For CNNs without local max pooling, global max, average, ROI and attention pooling underperform the baseline CNNs with global pooling. On the other hand, our attention-based atrous CNNs achieves the highest accuracies of .727 on subtask A and .732, .644, .622 on subtask B. Our model significantly outperforms the CNNs without local max pooling, which achieves accuracies of .604 on subtask A and .619, .467, .522 on subtask B (in a one-tailed z-test, $p < .001$ for subtask A and subtask B (device A and B), and $p < .05$ for device C in subtask B) . This result shows that the size of receptive field has a greater effect on the performance than a local max pooling operation. The atrous CNNs also fix

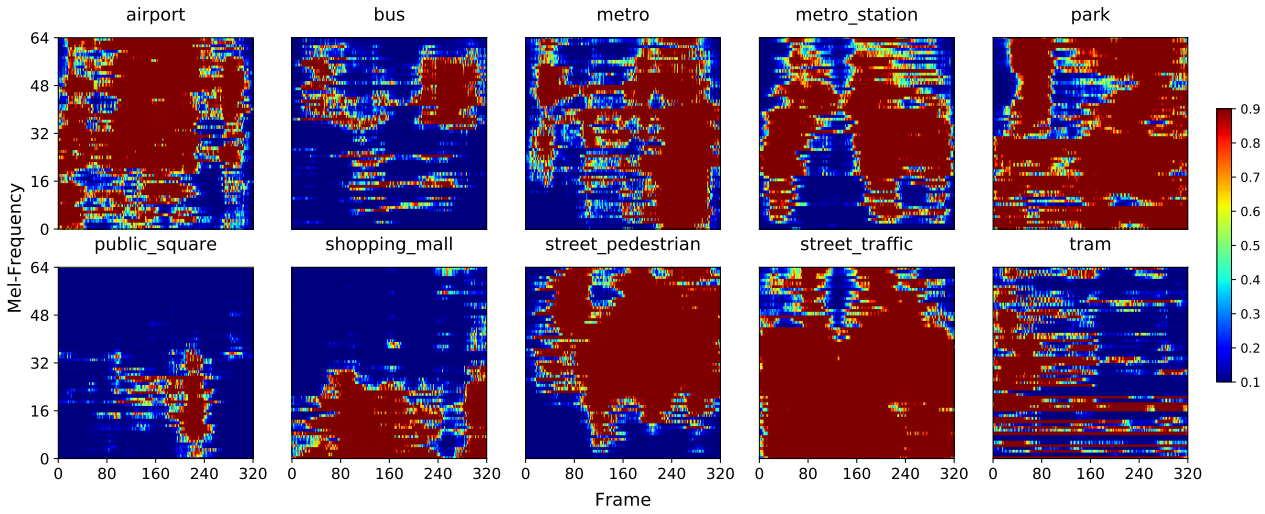

Fig. 3. Heat maps with a size of 64×320 are the visualisation of the attention matrix A in our attention-based atrous CNNs. The horizontal and vertical axes each represent the time frames and frequency bins.

Table 2. The class-wise accuracies of the result of attention-based atrous CNNs, which lead to the best results on two subtasks (SUBA on device *A*; SUBB on devices *A*, *B*, and *C*).

Accuracy	SUBA	SUBB		
Class	*A*	*A*	*B*	*C*
airport	.596	.740	.611	.389
bus	.777	.694	.667	.944
metro	.640	.816	.944	.556
metro_station	.757	.822	.667	.667
park	.843	.868	.778	.778
public_square	.593	.454	.500	.333
shopping_mall	.885	.681	.944	1.000
street_pedestrian	.522	.680	.444	.611
street_traffic	.894	.902	.833	.889
tram	.762	.663	.056	.056
Average	.727	.732	.644	.622

the resolution of the feature maps as 64×320, which can be visualised to observe the contributions of the time-frequency components in a feature map.

The class-wise accuracies are shown in Table 2. Our proposed model performs well for most classes on devices *B* and *C*, except *tram*. We think this might be caused by a lot of noise in recordings of *tram* by devices *B* and *C*.

4.4. Visualisation of the Feature Maps

The feature maps of the attention model are visualised in Fig. 3. For different acoustic scene classes, the contributions of each time-frequency unit are different. For example, *airport*, *park*, and *street traffic* mainly contain stationary background noise so that most time-frequency units have similar weight values. The temporal continuity at several fixed mel-frequency bins appears in the traffic environments, including *bus*, *metro*, and *tram*. The feature maps of *public square*, *shopping mall*, and *street pedestrian* indicate that some audio events like speech occurred.

5. CONCLUSIONS

This paper proposed attention-based atrous convolutional neural networks (CNNs) to visualise and understand acoustic scenes. Four dilated convolutional layers followed by a global attention pooling model were used to fix the size of feature maps for a visualisation. Our proposed model performed significantly better than the CNNs without dilation on the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 challenge tasks. Moreover, the time-frequency information in feature maps were visualised and analysed.

In future works, feature level attention models will be investigated to reach a deeper visualisation of CNNs. Further, CNNs followed by sequence to sequence learning methods and 3D CNNs will be considered to investigate the temporal information in acoustic scenes.

6. ACKNOWLEDGEMENT

This work was partially supported by the European Union's Horizon H2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 766287 (TAPAS), the EPSRC grant EP/N014111/1 "Making Sense of Sounds", and a Research Scholarship from the China Scholarship Council (CSC) No. 201406150082. We thank Judith Dineley for her proof-reading work.

7. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[2] G. Richard, T. Virtanen, J. P. Bello, N. Ono, and H. Glotin, "Introduction to the special section on sound scene and event analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1169–1171, May 2017.

[3] K. Qian, C. Janott, Z. Zhang, J. Deng, A. Baird, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Teaching machines on snoring: A benchmark on computer audition for snore sound excitation localisation," *Archives of Acoustics*, vol. 43, no. 3, pp. 465–475, Feb. 2018.

[4] S. Chu, S. Narayanan, C.-C. Kuo, and M. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. ICME*, Toronto, Canada, 2006, pp. 885–888.

[5] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 414–454, May 2013.

[6] Y. Xu, W. J. Li, and K. K. Lee, *Intelligent wearable interfaces*, John Wiley & Sons, 2008.

[7] K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, and B. Schuller, "Wavelets revisited for the classification of acoustic scenes," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 108–112.

[8] A. Vafeiadis, D. Kalatzis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Acoustic scene classification: From a hybrid classifier to deep learning," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 123–127.

[9] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. IJCNN*, Killarney, Ireland, 2015, pp. 1–7.

[10] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 113–117.

[11] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, May 2018.

[12] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. ISMIR*, New York, NY, 2016, pp. 805–811.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, Las Vegas, NV, 2016, pp. 2921–2929.

[14] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time-frequency segmentation from weakly labelled data," *arXiv preprint arXiv:1804.04715*, 2018.

[15] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association of Computational Linguistics*, vol. 4, no. 1, pp. 259–272, Jun. 2016.

[16] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. W. Schuller, "Attention-based convolutional neural networks for acoustic scene classification," in *Proc. DCASE*, Surrey, UK, 2018, 5 pages.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 12, pp. 2481–2495, Dec. 2017.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Boston, MA, 2015, pp. 3431–3440.

[20] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep Spectrum features," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3512–3516.

[21] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, "Learning image-based representations for heart sound classification," in *Proc. DH*, Lyon, France, 2018, pp. 143–147.

[22] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. DCASE Workshop*, Munich, Germany, 2017, pp. 17–21.

[23] L.-J. Li, H Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. NIPS*, Vancouver, Canada, 2010, pp. 1378–1386.

[24] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Proc. ICANN*, pp. 92–101. Thessaloniki, Greece, 2010.

[25] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. ISCA Speech Synthesis Workshop*, Sunnyvale, CA, 2016, pp. 125–125.

[26] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. ICLR*, Banff, Canada, 2014, no pagination.

[27] H. Caesar, J. Uijlings, and V. Ferrari, "Region-based semantic segmentation with end-to-end training," in *Proc. ECCV*, Amsterdam, Netherlands, 2016, pp. 381–397.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. DCASE Workshop*, Surrey, UK, 2018, 5 pages.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, 2015, no pagination.